适用于多维迫选测验的 IRT 计分模型

刘娟 [1] 郑蝉金 [2] 李云川 [1] 连旭 [1]

([1] 北京智鼎优源管理咨询有限公司，北京 100102) ([2] 华东师范大学，上海 200062)

摘 要 迫选(forced-choice, FC)测验由于可以控制传统李克特方法带来的反应偏差，被广泛应用于非认知测验中，而迫选测验的传统计分方式会产生自模式数据，这种数据由于不适合于个体间的比较，一直备受批评。近年来，多种迫选 IRT 模型的发展使研究者能够从迫选测验中获得接近常模性的数据，再次引起了研究者与实践人员对迫选 IRT 模型的兴趣。首先，依据所采纳的决策模型和题目反应模型对 6 种较为主流的迫选 IRT 模型进行分类和介绍。然后，从模型构建思路、参数估计方法两个角度对各模型进行比较与总结。其次，从参数不变性检验、计算机化自适应测验(computerized adaptive testing, CAT)和效度研究 3 个应用研究方面进行述评。最后提出未来研究可以在模型拓展、参数不变性检验、迫选 CAT 测验和效度研究 4 个方向深入。

关键词 迫选测验，自模式数据，TIRT，MUPP，GGUM-RANK

1 引言

心理测评可依据测量的内容分为认知测验和非认知测验。认知测验测量个体认知能力，如数值计算能力。这种测验通常具有标准答案，答对即得分，总分越高代表其相应的能力越高。非认知测验是了解个体的性格特点、价值观和态度倾向等方面最重要的方法之一，被广泛应用于临床心理诊断、职 [1]业生涯规划、人事决策中，有相当多的效度研究证明了性格对工作绩效有很好的预测效力 (SHL, 2018; Sitser et al., 2013; Hurtz & Donovan, 2000)。与认知测验不同的是，大部分非认知类的心理测评通常使用李克特形式的等级评定量表(rating scale)，其要求个体每次独立地评价一个题目(如，我是一个做事有条理性的人)，从最不符合我-1 到最符合我-5(5 级李克特)中选择与自己最接近的一项，答案没有对错之分。当在应聘、选拔等高利害的测评情境中使用此种题型的量表时，个体很容易有意地操控某些题目(如体现高责任心、乐观性的题目)的分数使自己看起来更符合组织期望，即使自己并不是这样的

收稿日期: 2021-07-06

通信作者: 郑蝉金, E-mail: chjzheng@dep.ecnu.edu.cn

人。这种可能的倾向被称为作假、装好，由此得到的测评结果便失去了对人才的区分效力，严重损害了测验的公平性。

为了消除或降低作假倾向的影响，通常会采用事前控制或事后控制技术(骆方,张厚粲, 2007)。事后控制技术包括嵌入作假识别量表、使用双因子模型控制作假因素(Brown et al., 2017; Hendy et al., 2021)、使用混合 Rasch 模型甄别作假反应模式人群(骆方,张厚粲, 2007)和基于历史数据建立决策树模型判别作假人群(Ziegler et al., 2012)等，其目的都是识别出作假数据，以避免依据作假数据做相关决策。这些方法均涉及的一个关键问题是如何保证较高的识别准确率，因为误将诚实的个体判为造假是非常不可取的，而相对于处理已经受到污染的数据，事前控制技术旨在阻止个体在答题前或答题中作假以获得无污染数据，这类技术包括警告、假渠道技术(bogus pipeline)和迫选测验。警告是其中最容易操作的方法，分为作假识别警告和后果警告，前者为告知个体可以识别到他们回答中的任何不诚实行为，后者为告知个体不如实作答会带来什么后果，Dwight 和 Donovan(2003)的元分析结果表明后果警告才能起到抑制作假的作用，而在他们后来的研究中进一步指出了两种警告方式一起使用才能产生统计学上有意义的结果。另外，当个体被警告时，潜在的作假者可能会决定在作答时不那么极端，或者为了看起来更诚实而选择一些“错误”的答案，即警告可能会诱发更加老练的作假，那么警告并没有实质性的改变个体作假的机制。警告还可能会造成被试在测验过程中焦虑程度提升等一些负面影响，因此只警告那些表现出作假趋势的被试被认为是更好的解决办法，而决策树模型可用于决定何时警告作假者(Ziegler et al., 2012)。假渠道技术是通过故意引导个体以为其在进行测谎实验(实则是在进行真实测评)，以迫使其做出最真实的反应，即给真实测评套上了一个测谎目的的壳子。但这种技术是在欺骗个体的基础上实现的，有违伦理道德，因此备受谴责(Aguinis & Handelsman, 1997)。

迫选测验要求个体在一组称许性水平相似的题目中强制选择最符合自己和最不符合自己的两项，或对题目进行偏好排序，个体无法对所有题目都给予积极的选择。由于题目的称许性相似，没有一个题目比其他题目更可取，那么个体根据社会称许性做选择/作假的可能性就会降低。相比李克特式量表，个体更不易在迫选测验上作假(Saville & Willson, 1991; Jackson et al., 2000; Wetzel et al., 2020)。强制选择的作答形式也消除了李克特式量表其他的一些潜在作答反应偏差(responses biases)，如光环效应，趋中倾向，极端倾向，默许(总是选择同意或不同意)等。另外，迫选测验形式能有效降低分数在社会称许方向的膨胀性(Cao & Drasgow, 2019)，也没有明显降低个体的作答积极性(Sass et al., 2020)或给个体带来情绪或认知上的不利影响(Zhang et al., 2020)。Bartram(2007)的一项元分析结果表明，相比李克特式评

定量表，由迫选测验获得的评估结果对工作绩效的预测效度可以提升 50%。但迫选测验的传统计分方式会产生自模式数据(ipsative data)，分数的高低体现了个体在各个维度上内部自比的排序结果，这种数据形态的特殊性限制了迫选测验在个体间比较场景（如人才选拔）中的应用与发展。近十几年来，几种迫选测量模型的发展使研究者能够从迫选测验中获得接近常模性的潜在特质估计结果，克服了自模式数据问题后的迫选测验似乎成为了更有应用潜力的抗作假技术。

本文旨在系统地介绍迫选测验的题目类型、特点及传统计分方式和自模式数据的弊端，然后从题目反应模型和决策模型两个方向，介绍与评价 6 种迫选 IRT 模型，其次从模型构建思路、参数估计方法和应用研究现状几个方面对比分析 6 种模型，最后从迫选模型实践的角度提出 4 个未来研究的展望方向：模型拓展研究、参数不变性研究、迫选 CAT 研究和效度研究。

2　迫选测验设计与传统计分方式

迫选测验通常由测量不同维度的数个迫选题块(item block)组成。题块内由固定数量的来自不同或相同维度的、社会称许性水平相似的题目/描述(item/statements)组成，题目/描述即为维度（也即潜在特质）的外显指标。同一题块的题目通常分别测量不同维度，因此也被称为多维迫选题(multidimensional forced-choice, MFC)。

2.1　迫选测验设计

根据 Hontangas 等(2015)的分类，迫选题块有 3 种常见的形式：PICK、RANK 和 MOLE。这种分类主要体现在指导语类型上。PICK(表 1)要求个体从题块中选择最符合自己的一项。RANK(表 2)要求个体对题目进行从最符合到最不符合的完全排序。MOLE(表 3)要求个体分别选择出最符合自己(MOst)和最不符合自己(LEast)的一项。超过 3 个题目的 MOLE 题型也称部分排序题(partial rankings)。

表 1　PICK 题型

指导语：从以下两个描述中选择最符合自己的一项	
题块	最符合
A 寻找事物的不足	√

B 探索陌生的领域

表 2 RANK 题型

指导语：对以下描述进行排序	
题块	排序
A 寻找事物的不足	3
B 探索陌生的领域	1
C 基于数据分析做决定	2

表 3 MOLE 题型

指导语：从以下描述中选择最符合自己和最不符合自己的一项		
题块	最符合	最不符合
A 寻找事物的不足		
B 探索陌生的领域	√	
C 基于数据分析做决定		
D 做注重精确性的工作		√

题块大小即题块内包含多少个题目/描述选项，2~4 个题目的题块大小是最为常见的。为节省篇幅，在后文中将结合指导语类型和题块大小对迫选题型做简称，如称 3 题目题块的 RANK 题型为 RANK-3。题块大小会影响个体选择任务负荷的高低，PICK-2 仅需个体将 2 个题目对比一次即可，题目越多，个体需要进行题目间对比的次数越多，使用大题块会增加选择任务的认知复杂性，可能对受教育程度较低或阅读能力较差的人有不利影响(Brown, 2016)。目前已有的迫选测验中较为常用的题块类型有：PICK-2 (Oswald et al., 2015)、RANK-3 (连旭 等, 2014; SHL, 2018)、MOLE-4(SHL, 1997)。其中 RANK-3 既没有 MOLE-4 的高认知负荷，也比 PICK-2 更加高效，且提供的信息量也最大(Hontangas et al., 2015; Joo et al., 2018)。

另外还有 Q 分类(Q-Sort)(Block, 1963)这种特殊的迫选题型，它是将问卷中所有的题目(如超过 30 个题目)组合为一个大型题块一起呈现给个体，然后要求个体逐步地将每个题目分配到少数几个偏好等级中，如先从所有题目中选择出最符合自己的几个题目，然后从剩下的题目中选择最不符合的几个，直至完成所有题目的分类。这种方法需要个体一次处理大量描述，因此适用于词汇型的题目(Brown, 2016)。

迫选测验组卷时，需考虑的首要原则是题目称许性的匹配，这是保证测验具有抗作假效力的关键步骤，然后才是题块大小、指导语等外显因素。通常会计算题目称许性的平均绝对差值来衡量匹配程度，差值越大代表越不匹配，然而这种仅用均值判断的方式会忽略不同评价者对同一题目称许性评价的差异。Pavlov 等人(2021)提出了一种替代性指标：IIA(Inter-

item agreement)指数，该指数将 BP 指数和 AC 指数(Gwet, 2014)纳入到题目称许性的匹配中，可更好地匹配那些原本在称许性均值上没有差异的题目。实践人员可借助 R(R Core Team, 2021)包 autoFC (Li et al., 2021)计算 IIA 指数并进行自动组卷。

2.2 传统计分方式与自模式数据

2.2.1 传统计分方式

通常，迫选测验的传统计分方式是将每个题块中被选为最符合或排序最高的题目计 1 分，最不符合或排序最低的题目计-1 分，未选择或中间等级的题目计 0 分，最后将各维度下题目分数进行累加得到维度总分。

题目的描述方向将影响各维度题目计分的方式，负向描述的题目在计分时需乘以-1 进行分值的转换，如负向描述(如：我时常预期消极的结果)被选为最符合时需计为-1 分。负向描述的称许性通常很低，所以很难匹配不同计分方向题目的称许性。如果将称许性相差较大的正负向题目放在一个题块里，个体很容易选择正向题目为更符合。尤其是在高利害情境中，几乎所有被试都会选择看起来更积极的选项(Bürkner et al., 2019)，进而会产生测量精度问题，也会使测验丧失抗作假的作用，因而实际应用中很少使用混合计分型的题块。

2.2.2 自模式数据及其问题

以表 3 的 MOLE-4 题型为例，无论个体如何选择，其在每个题块上所选出最符合和最不符合的题目都将分别计为 1 和-1 分，那么各个题块内题目的得分和均为 0，进而在整个测验上的总分也为 0。由此可见，各个维度的得分是互相依赖的，有高分维度则必然存在低分维度，不会出现所有维度得分同高或同低的情况，这种数据则为自模式数据。与之相对的是常模性数据(normative data)，如李克特式量表的数据，不同个体对每个题目的评定是互相独立的，评价分数互不影响，因此测验的总分是不固定的。自模式数据内部的分数依赖性违反了经典测验理论的基本假设之一，即误差方差的独立性，这对迫选测验分数的统计分析和解释都有影响(Baron, 1996)，如信度分析、方差分析、回归分析等，它会增加犯 I 类错误的概率，同时也会影响统计检验力(王珊 等, 2014)。同时，自模式数据对维度关系的扭曲会污染测验的结构效度与效标关联效度(Brown & Maydeu-Olivares, 2013)，并无法进行因子分析(Closs, 1996)。最后，将自模式数据做常模化的分数解释，进行个体间对比是不妥的，这可

能会扭曲个体的真实情况，如在兴趣测验中，自比结果仅代表个体内部的倾向性排序，Closs 认为直接进行人群间比较会严重高估或低估个体真实的兴趣特征。

迫选测验所测的维度数量及维度间的关系对数据自模程度的影响较大。Bartram(1996)的研究表明当维度数量低于 10 个时或者维度间的相关性达到 0.3 及以上时，自模式分数结果将不可靠，且信度会随维度个数的降低和维度间相关性的提高而大幅度降低。Clemans(1966)也指出低维度数量的迫选测验意味着更严重的自模式数据问题。Baron(1996)指出如果真实分数均匀地分布在平均值周围，那么自模式分数与常模性分数就会相似，反之如果多数维度高于或低于平均值，自模式分数与常模性分数则有很大区别，但这种区别会随着测验所测的维度数量的增多而下降，因为人群中出现多个维度分数同高或同低的可能性会大幅下降。相似的，当维度间的相关关系均为高正相关或高负相关时，出现分数同高或同低的可能性也会变高，当维度间的相关关系有正有负时，出现同高或同低这种高度偏态的维度特征的可能性就低了很多。Saville 和 Willson(1991)的研究也证明当维度数量超过 30 个且维度内部相关性较低时，由自模式数据计算的测验信度达到了可接受范围，且维度的性状恢复性与常模性数据相似，此时使用常模化的自模式数据进行分数解释和个体间的比较是可行的。因此，增加测验的维度数量是抵抗自模式数据特点较为有效的传统做法之一，但也只是折中的办法。

综上，自模式数据的诸多问题限制了迫选测验的应用，虽然可通过增加维度等方法抵抗自模性问题，但可以看到，传统计分方式是把个体对题目的排序结果当作对其的绝对评分，它并未体现个体比较决策的心理过程，应用在迫选测验上是不恰当的。要解决自模式数据的问题，需要从根本上跳出传统计分方式，采用现代测量模型来反映个体在回答迫选题目时的决策过程(Brown & Maydeu-Olivares, 2013)，从外显的比较结果中获得影响决策过程背后的潜在特质分数，从而实现恢复个体分数的常模性。

3 用于迫选测验的 IRT 计分模型

在过去的十几年间，众多适用于迫选测验的 IRT 计分模型被开发出来以建立外显作答与潜在特质的关系，从而获得具有常模性特点的潜在特质分数进而实现个体间分数的比较。其中被研究与应用最为广泛的模型之一是由 Brown(2011)提出的瑟斯顿 IRT 模型(Thurstonian Item Response Theory, TIRT)，由 Stark 等人(2005)提出的 MUPP(Multi-Unidimensional Pairwise Preferences)框架也因其灵活性在近几年引起了较多研究者的关注，并发展出了 2 个新模型

(Morillo et al., 2016; P. Lee et al., 2019)。另外还有 Wang 等(2017)开发的 Rasch 自模模型(Rasch ipsative model, RIM)，H. Lee 和 Smith (2020a)基于贝叶斯题组模型(Bayesian testlet model) (Bradlow, Wainer, & Wang, 1999)提出的贝叶斯随机题块模型(Bayesian random block item response theory, BRB-IRT)。这些模型均包含三个层面的内容：迫选题型、题目反应模型、决策模型。模型之间的本质区别在于所假设的题目反应模式(Morillo et al., 2016)和采用的决策模型类型 (Brown, 2016)。题目反应模式反映的是题目反应强度和所测维度之间的关系，决策模型类型反映的则是个体在题目间做出选择的过程。题型和决策模型共同决定了模型的基础框架，决策模型在外显作答与题目反应强度之间起到了桥梁作用，并进一步由题目反应模型链接到个体的潜在特质水平，最终形成整体的迫选分析模型。本文将首先厘清不同题目反应模式和决策模型类型，再依据这两种概念类型对上述模型进行分类和系统介绍，最后从模型构建思路、参数估计方法与应用研究现状 3 个方面进行模型比较。

3.1 题目反应模式

题目是特质的外显测量指标，题目与潜在特质之间的关系需使用测量模型进行链接。在人格测验中，不同测量模型依据其所假设的个体对题目的反应过程，可划分为优势模型(Dominance Models)和展开模型(Unfolding Models)两大类。优势模型假定个体被评估的特质水平越高，其会以越高的概率对相应题目做出正面回答，Rasch 模型、2PL 模型(Two-Parameter Logistic Model，2PLM)等均假设个体对题目的回答遵循优势反应模式。展开模型假定个体正面回答的概率与题目和被评估的特质水平位置的接近程度直接相关。如题目“我喜欢和朋友在咖啡馆里安静地聊天”，太过内向的个体会因为不喜欢公共场所而选择不同意，而极端外向的个体因为喜欢更加刺激的环境而选择不同意 (Drasgow et al., 2010)，处于中间水平的个体更倾向于同意，其项目反应函数曲线为单峰钟型，即个体的特质水平与题目位置越接近，其正面回答的概率越高。展开模型的代表模型为广义等级展开模型(Generalized Graded Unfolding Model, GGUM) (Roberts et al., 2000)。

到底哪种模型更能反映出个体在作答非认知类题目时的反应特点，至今仍未有定论(王珊 等, 2014; Morillo et al., 2016; Hontangas et al., 2016)。一些模拟和实证研究(Chernyshenko et al., 2001; Tay et al., 2011)支持展开模型，特别是针对态度类特质的测量，展开反应题目的表现与优势反应题目一样好或更好。展开模型被认为更灵活，因为当题目的位置参数在末端时，它可以等同于优势模型。然而研究表明这种优越性在实践中并非普遍存在，与优势反应

题目组成的量表相比，完全由展开反应题目组成的量表的心理测量学特性大为逊色，包括较低的信度和较低的效标关联性(Huang & Mead, 2014)。此外，由于展开模型对负向题的评分无法直接反向转换，估计结果可能不如优势反应题目准确(Brown & Maydeu-Olivares, 2010)。从模型复杂度上来说，优势模型一般比展开模型更节俭、有更少的参数，通常情况下除非有明确的证据证明复杂模型的优势，否则应首先考虑更节俭的模型(Oswald & Schell, 2010)。另外，展开反应题目更加难以编写，题目所反映的确切含义也难以界定。更多关于优势还是展开模型的讨论可参考 Drasgow et al. (2010)。

题目反应模式是题目层面的特点而非特质特点，与迫选题型无关。在将单个题目组合为迫选题块时，可使用任何反应模式的题目，因为它们都能测量同样的潜在特质，潜在特质的分布对于同一批人群来说是不变的。在实际应用中，需要研究者结合题目特点或数据特点，选择优势或展开模型中的一种作为题目与潜在特质间的测量模型，尚未看到在同一测验中混用两种模型的情境。

3.2 决策理论

迫选测验要求个体对一组题目进行比较判断进而决策产生答案，而非对每个题目进行独立的评价，而个体对题目的绝对评价是衡量其特质水平的基础。基于 Brown(2016)的观点，个体对一组题目进行比较判断的基础是其在每个被比较题目上的绝对评价水平，对迫选数据的建模需要依托于合适的决策理论来阐释决策结果(外显作答)与绝对评价之间的关系，进而评估个体的潜在特质水平。目前已被用于迫选数据建模的决策理论主要有两类，第一类是最古老和被使用最广泛的瑟斯顿比较判断法则(Thurstone’s Law of Comparative Judgment)(Thurstone, 1927)，第二类是 Luce 选择公理(Luce & Duncan, 1959)和布拉德利-特里模型(Bradley–Terry Model) (Bradley & Terry, 1952)，后者是前者的特例情况(Brown, 2016)。

3.2.1 瑟斯顿比较判断法则

Thurstone(1927)以效用(utility)来表示个体对每个题目的反应倾向性大小。效用是一个潜在变量，可以被认为是一个题目在个体上的心理价值。Thurstone 认为个体对题目的权衡考虑实质是效用值的衡量。以 y_{ij} 代表个体比较题目 i 和 j 后的外显结果，$y_{ij}=1$ 代表选择了题目 i 为最符合，否则 $y_{ij}=0$。以 t_i 表示个体在题目 i 的效用值，$t_i > t_j$ 表示个体在题目 i

上的效用高于题目 j，更倾向于选择题目 i 而非题目 j，那么以 $y_{ij}^{*}=t_i-t_j$ 表示题目 i 与题目 j 的效用差值，效用与外显作答的关系可整理如下：

$$y_{ij}=\begin{cases}1, & y_{ij}^{*}\geq 0\\ 0, & y_{ij}^{*}<0\end{cases} \tag{1}$$

在应用到迫选模型建模时，不同个体在题目 i 上的效用差异可被分为系统与随机两部分，系统部分 $f(\theta_a)$ 可以是与个体潜在特质水平相关的反应函数，随机部分则为随机误差 ε_i，Thurstone 假设其在不同题目间相互独立且服从正态分布。因此，效用与潜在特质之间的关系可以下式表示：

$$t_i=f(\theta_a)+\varepsilon_i \tag{2}$$

其中 θ_a 为个体在题目 i 所测量的潜在特质 a 上的水平。

3. 2. 2 Luce 选择公理

Luce(1959,1977) 拓展了适用于二元选择情境(binary choice)的布拉德利-特里模型(Bradley & Terry, 1952)，其以 v_i 代表某个体与题目 i 相关的反应强度，将由所有备选题目组成的集合称为 S，那么从 S 中选择 i 的概率 $P(i[S])$ 与 v_i 成正比：

$$P(i[S])=\frac{v_i}{\sum_{k\ in\ S} v_k} \tag{3}$$

Luce 将对一组题目的排序过程描述为互相独立的一系列做最佳选择的步骤：从题目集合 S 中先选择最符合自己的题目 i，再从剩余的 $S-1$ 集合中选取第 2 个最符合的题目 j，然后从剩余的 $S-2$ 集合中选取第 3 个题目，直到完成最后两个题目的选择，从而实现对所有备选题目的排序(Hontangas et al., 2015)。排序结果的概率则为各步骤概率的连乘。假设 S 集合包含 i、j、k 三个题目，那么排序结果为 $i>j>k$ 的概率为：

$$P(ijk)=P(i[ijk])\times P(j[jk]) \tag{4}$$

其中，$P(i[ijk])=\frac{v_i}{v_i+v_j+v_k}$，代表从 i、j、k 中选择 i 的概率；$P(j[jk])=\frac{v_j}{v_j+v_k}$，代表从 j、k 中选择 j 的概率。

当集合 S 中仅有 i、j 两个题目时，Luce 选择公理的应用即为布拉德利-特里模型：

$$P(i[ij]) = \frac{v_i}{v_i + v_j} \tag{5}$$

将此决策模型应用到迫选模型建模中时，v_i 可由与潜在特质有关的项目反应函数得出。

其他类型的决策理论还有，如 Coombs 的展开偏好模型(Coombs's Unfolding Preference Model)、Andrich 的强制赞同模型(Andrich's Forced Endorsement Model)。前者是瑟斯顿比较判断法则的一个特例，后者简化后与布拉德利-特里模型等价，具体可参考 Brown(2016)。

3.3 TIRT 模型

TIRT 是 Brown (2011)基于瑟斯顿比较判断法则提出的一种适用于优势反应题目的模型，它适用于 PICK-2，RANK 和 MOLE 题型的迫选测验，题块中的题目可以来自同一维度也可属于不同维度。TIRT 假设个体选择或者排序的心理过程是依次地对一个题块内 n 个题目进行了独立地两两比较判断，这个过程产生了 $\tilde{n} = n(n-1)/2$ 个比较结果，在对数据进行建模前，需要对作答进行二元编码(binary coding)以获得两两题目的比较结果。以一个 RANK-3 题块为例，题块内的题目为 $\{i, j, k\}$，假设个体的选择结果为 $i > k > j$，编码结果则为 $\{i, j\} = 1$，$\{i, k\} = 1$，$\{j, k\} = 0$，代表 $i > j$，$i > k$ 和 $j < k$，TIRT 是在拆分后的二元数据上构建的概率模型。

在 TIRT 中，效用与题目所测的潜在特质之间是线性关系，且假设每个题目的效用均只在一个潜在特质上有载荷，即题目是单维性的。假设题目 i 测量了特质 a，结合公式(2)，效用与特质 a 的关系可以表示为：

$$t_i = \mu_i + \lambda_i \theta_a + \varepsilon_i \tag{6}$$

其中 μ_i 为潜在效用 t_i 的均值，λ_i 为题目 i 在潜在特质 θ_a 上的因子载荷，ε_i 为服从正态分布的误差，θ_a 被假设服从多元正态分布。实际中，研究者通常更关注 θ_a 的大小，而非效用值，因此需要通过公式(1)建立起效用值、潜在特质和外显作答之间的联系。假设题目 j 测量潜在特质 b，将公式(6)代入 $y_{ij}^* = t_i - t_j$ 中可得：

$$y_{ij}^* = \mu_i + \lambda_i \theta_a + \varepsilon_i - \mu_j - \lambda_j \theta_b - \varepsilon_j \tag{7}$$

使 $\mu_i - \mu_j = \gamma_{ij}$，且假设潜在特质、误差均服从正态分布，那么 y_{ij}^* 也服从正态分布。基

于 Thurstone 对 ε_i 的正态性假设要求采用正态肩型模型(Normal Ogive Model) 为链接函数，那么对于每个二元结果，个体选择题目 i 而非 j 的条件概率为：

$$P\left(i > j \mid \theta_a, \theta_b\right) = \Phi_N\left(\frac{\gamma_{ij} + \lambda_i\theta_a - \lambda_j\theta_b}{\sqrt{\psi_i^2 + \psi_j^2}}\right) \tag{8}$$

其中 ε_i 和 ε_j 的方差为 ψ_i^2、ψ_j^2，那么差值的方差为 $\psi_i^2 + \psi_j^2$，Φ_N 代表累积正态分布函数。经过二元编码后的数据两两之间存在共同题目，如 $\{i, j\}$ 与 $\{i, k\}$ 均包含对 i 题目的判断，因此它们之间的协方差将被设定为共享成分 i 的方差，以解释其之间的相互依赖性。

$$\operatorname{cov}\left(\varepsilon_i - \varepsilon_j, \varepsilon_i - \varepsilon_k\right) = \operatorname{cov}\left(\varepsilon_i, \varepsilon_i\right) = \psi_i^2 \tag{9}$$

所以公式(8)是一个特殊的二维正态肩型 IRT 模型，潜在特质 θ 越高，则个体选择题目 i 而非 j 的概率也越高，遵循优势反应模式。

TIRT 发展至今，有众多研究者通过模拟与实证研究探索了其在多种条件下的适用性(Bürkner et al., 2019; Brown & Maydeu-Olivares, 2013; Schulte et al., 2021; 李辉 等, 2017; 连旭 等, 2014)。这些研究一方面证明了 TIRT 确实在一定程度上克服了传统计分下的自模性问题，相比传统计分具有测量精度的提升，也更接近李克特式单一刺激量表的结果(Joubert et al., 2015)；另一方面也指出了 TIRT 若要显示出比传统计分优良的性质，需对测验设计有较多限制。如 TIRT 在低维度数情境中使用时，其潜在特质的良好恢复性建立在测验包含一定比例的混合计分型题块的基础上(Brown, 2011)。Schulte 等(2021)的研究也指出在维度数量低于 10 时，如果所有题目同为正向题，即使是在高因子载荷情况下，测验的信度也会急剧下降。不过与传统计分的相关研究相似，在高维度情境下(维度数量高于 30)，即使不使用混合计分型题块，TIRT 对潜在特质分数及特质间关系的恢复性也非常准确(Schulte et al., 2021; Bürkner et al., 2019)。最后需注意的是，使用混合计分型题块可能存在以下几个问题(Bürkner et al., 2019; Morillo et al., 2016)：1、增加个体的认知负荷；2、反向描述可能会带来较大的方法论变异，可能会组成一个独立的方法因子，进而会影响题目的协方差矩阵；3、可能会损害使用迫选题型来控制作假的效力，进而导致对使用迫选测验意义性的质疑。

3.4 MUPP 框架及衍生模型

3.4.1 MUPP 框架与 MUPP-GGUM 模型

Stark(2005)提出了适用于配对迫选题型(PICK-2)的 MUPP 框架，该框架对后来迫选模型的发展起到了极大的促进作用(Brown & Maydeu-Olivares, 2013)。在 MUPP 中，假设一个题块包含题目 i 和 j，并分别测量潜在特质 θ_a 和 θ_b，以 $P(i)$ 代表个体对题目 i 的接受概率，$Q(i)$ 代表对题目 i 的拒绝概率，且 $Q(i)=1-P(i)$，那么个体选择题目 i 为最符合的反应概率 $P(i>j\,|\,\theta_a,\theta_b)$ 为：

$$P(i>j\,|\,\theta_a,\theta_b)=\frac{P(i)Q(j)}{P(i)Q(j)+Q(i)P(j)} \tag{10}$$

MUPP 假设个体对每个题目的评定是独立的，且题目是单维性的，题块内的题目可来自相同或不同维度，因此被称之为多重-单维配对选择模型(Multi-Unidimensional Pairwise Preference, MUPP)。MUPP 反映了个体决策结果的概率与单个题目倾向程度的关系，如果把 $P(i)Q(j)$ 的联合概率记为 v_i，表示为与题目 i 相关的反应强度，则其与公式(5)是等价的，因此 MUPP 采用的决策模型可归类为布拉德利-特里模型(Brown, 2016)。

在题目反应模式的选择上，Stark(2005)假设题目服从展开反应模式，并使用 GGUM 的二元计分版本计算单个题目的反应概率，即公式(10)中的 $P(i)$ 与 $Q(j)$，因此该迫选模型被称为 MUPP-GGUM 模型。为方便研究者应用此模型，Stark(2002,2005)提出了组建 PICK-2 题型迫选测验的建议流程：

1、 为每个所要测量的维度出大量的题目描述(建议 3 倍于目标题量)；

2、 将题目以 1~4 级量表或 1~5 级量表进行施测(各维度约 1000 人左右的被试)；

3、 分维度估计题目的参数，并进行单维性检验；

4、 对题目进行社会称许性量表的评定，取人群平均值作为题目的称许性水平；

5、 通过前四步完成迫选题库的搭建后，就可以将称许性等级相似且测量不同特质的题目进行配对组卷，以减少个体依据称许偏好作答。为确定潜在特质分数的尺度，需要包含一定比例的同维度题对；

6、 投放迫选测验施测；

7、 使用 MUPP-GGUM 模型对个体进行特质的估计。

MUPP-GGUM 模型是使用最为久远和广泛的迫选模型之一，不仅在开发流程上有规范的流程指导，也是最先被应用到计算机化自适应测验开发的迫选模型，并应用在了美国军队选拔的多个人格测验中(Stark et al., 2012; Stark et al., 2014)。在 MUPP-GGUM 之后，有众多基于 MUPP 框架的衍生模型被开发出来以适应多种迫选题型，同时在 Stark 等人(2012)自适应算法的基础上，迫选自适应测验研究也开始蓬勃发展。

3.4.2 MUPP-2PL 模型

Morillo 等(2016)认为优势反应模式的题目也同样适用于非认知类测验，并且在题目编写难度、模型节俭度方面要优于展开反应模式的题目。因此在 MUPP 框架的基础上，Morillo 等人将公式(10)中计算 $P(i)$ 与 $Q(j)$ 的项目反应函数替换为了经典优势反应模型 2PLM，并称之为 MUPP-2PL 模型。依据此模型，个体选择题目 i 而非题目 j 的概率为：

$$P(i > j \mid \theta_a, \theta_b) = \Phi_L\left(a_i\theta_a - a_j\theta_b + d_{block}\right) = \frac{1}{1 + \exp\left[-\left(a_i\theta_a - a_j\theta_b + d_{block}\right)\right]} \quad (11)$$

其中，Φ_L 代表逻辑斯蒂克函数(logistic function)，a_i 和 a_j 代表题目的区分度参数，θ_a 和 θ_b 分别代表题目 i 和 j 所测量的潜在特质，d_{block} 为截距参数，其由 2PLM 中的 a、b 参数合并得到 $\left(d_{block} = a_i b_i - a_j b_j\right)$，但单个题目的 b 参数是无法被识别的。

MUPP-2PL 并不是唯一一个 MUPP 框架在优势反应模型下的应用，Usami 等(2016)也在 MUPP 中使用了 2PLM 来计算单个题目的接受程度，但其与 Stark(2005)一样采用了预标定的题目参数来估计能力。虽然使用基于单维模型进行参数标定的方法在算法和题库管理上较为简便，但从应用角度出发，人格类测验通常没有正确答案，在题目保密性的需求上并不突出，一般无需配置大型题库用于组合平行试卷，一套优秀的测验便足够，那么此时基于迫选作答数据来估计题目参数显然更符合真实情境(P. Lee et al., 2019)。同时 Stark 的这种方法除了忽略了题目参数跨测验情境的变异性之外，在估计个体潜在特质时也忽略了题目参数的估计误差。因此 Morillo 等(2016)基于贝叶斯框架，采用马尔科夫链蒙特卡洛采样(Markov chain Monte Carlo, MCMC)算法对题目参数和被试参数进行联合估计，实现了基于迫选作答数据来估计 MUPP-2PL 的所有参数。Morillo 等发现题目参数、能力参数和特质间的关系恢复性

均受到测验长度的影响，即测验越长估计结果越准确。另外，样本量是影响题目参数估计准确性的重要因素，该方法对 d_{block} 参数的估计相比 a 参数更加准确。最后，Morillo 等在实证研究中发现 MUPP-2PL 对部分特质之间关系的估计结果与前人研究有较大差异，但此差异来源是作答人群还是测验情境的改变尚不得而知。

3.4.3 GGUM-RANK 模型

MUPP-GGUM 与 MUPP-2PL 均只适用于 PICK-2 题型，Hontangas 等(2015)将 MUPP 框架依据 Luce 选择公理进行了拓展，使之能够适用于 PICK、RANK 和 MOLE 多种迫选题型。首先假设有 i 、 j 、k 共 3 个题目组合为一个题块，当为 PICK-3 题型时，那么基于 Luce 选择公理，个体从集合 $[ijk]$ 中选择 i 的概率 $P(i[ijk])$ 为：

$$P(i[ijk]) = \frac{P(i)Q(j)Q(k)}{P(i)Q(j)Q(k)+Q(i)P(j)Q(k)+Q(i)Q(j)P(k)} \tag{12}$$

而对 RANK 题型进行拓展的逻辑是，假设个体对题目的排序过程实则是对题目进行了一系列的 PICK，以 RANK-3 题型为例，假设一个个体的排序结果为 $i>k>j$ ，那么 $P(ikj)$ 则为：

$$P(ikj) = P(i[ijk]) \times P(k[jk]) \tag{13}$$

其中 $P(i[ijk])$ 由公式(12)得出，同理可得 $P(k[jk])$。

最后对 MOLE 题型进行拓展，以 MOLE-4 为例（增加题目 l ），未被选择的两个题目的排序无法确定，因此合并 2 种可能的排序作为此题型选择结果的概率。以 $P(i{*}{*}k)$ 表示被试选择了 i 和 k 作为最符合自己和最不符合自己的题目时的概率，那么：

$$P(i{*}{*}k) = P(ijlk) + P(iljk) \tag{14}$$

其中， $P(ijlk)$ 和 $P(iljk)$ 可基于公式(13)的逻辑计算。

以上基于 Luce 选择公理对 MUPP 的拓展，使得 PICK、RANK 和 MOLE 题型的判断逻辑被整合到一个框架内，形成嵌套关系，极大地拓宽了 MUPP 的应用范围。P. Lee 等(2019)则基于以上对 RANK 模型的拓展思路，开发了适用于 RANK-3 题型的 GGUM-RANK 模型（即公式 13 中的 $P(i)$ 由 GGUM 计算），并采用 MCMC 联合估计算法对题目参数和能力

参数进行估计。Joo 等(2018)开发了此模型的两种信息量指标：OII(Overall item information)和 OTI(overall test information)。OII 为一个题块的信息量，OTI 为测验中所有题块的信息量的累加和，即测验整体信息量。这两种信息量指标可为测验的组卷提供直接参考，而在挑选相似 OII 的题块时，Joo 等给出了一种绘制条件 OII 图形的方法，使研究者可以进一步比较和选择能够在目标能力区间内提供最大信息量的题块，而信息量指标的开发也为 GGUM-RANK 实现 CAT 打下了基础(Joo et al., 2020)。

3.5 RIM 模型

Wang 等(2017)认为通过迫选测验识别个体的潜在特质的绝对水平是不现实的，并指出通过 TIRT 获得的特质分数不能用于个体内和个体间的比较。因此，Wang 等提出了 RIM 模型，旨在获得用于个体内部比较的分数，而非像 TIRT 或 MUPP 族模型期望获得潜在特质的绝对分数，其使用 Rasch 模型作为项目反应函数，因此 RIM 适合优势反应模式的题目。与 TIRT 模型一样，RIM 的决策模型为 Thurstone(1927)的比较判断法则，个体对题目的比较实则衡量的是特质分数与题目效用值。在 RIM 模型中，个体选择题目 i 而非题目 j 的概率为：

$$P\left(i > j \mid \theta_a, \theta_b\right) = \Phi_L\left(\theta_a + \mu_i - \theta_b - \mu_j\right) \tag{15}$$

其中，θ_a 与 θ_b 为题目 i 和 j 所测量的潜在特质，μ_i 与 μ_j 为题目的效用值。

其在对潜在特质 θ 估计时，个体内部在所有所测特质上的分数和将被固定为 0，因此只有 $D-1$ 个潜在特质被自由估计：

$$\sum_{d=1}^{D} \theta_d = 0 \tag{16}$$

其中 d 代表维度，D 代表维度数量。此时 θ 的大小意味着心理特质的分化程度，如果 θ 接近 0 意味着更低的分化程度。所以特质的 θ 值实则是比自模计分更加精细的内部排序结果。$\theta_a - \theta_b$ 代表特质 a 与特质 b 之间的排序的相对差异，如果个体 m 和个体 n 在特质 a 与特质 b 上的差值的绝对值为 $ABS(\theta_{am} - \theta_{bm}) > ABS(\theta_{an} - \theta_{bn})$，那么个体 m 在特质 a 和 b 上的分化程度大于个体 n。在对模型的参数估计上，Wang 等建议当维度数量低于 4 个时可采用 MMLE(Marginal Maximum Likelihood Estimation)算法，高维数量时更适合用 MCMC 方法。

Wang 等(2016)拓展了 RIM，使之适用于 RANK 题型，形成了 ELIRT(exploded logit IRT)和 GLIRT(generalized logit IRT)两种迫选模型。其中 ELIRT 的拓展思路与 Hontangas 等(2015)

对 RANK 的拓展逻辑一致。GLIRT 的拓展思路是对每个题块的可能作答模式进行枚举，依次写出每种作答模式的反应函数，并限定所有可能的作答模式的概率和为 1，来实现对个体作答模式概率模型的构建，具体可参考 Chen 等(2020)。当用于配对迫选题型时，ELIRT 和 GLIRT 均等价于 RIM。两种拓展模型的模拟研究结果非常相似，研究者可自由选择其中之一。

3.6 BRB-IRT 模型

H. Lee 和 Smith(2020a)选择了贝叶斯题组模型(Bradlow et al., 1999)作为基础模型，通过在 MUPP-2PL 的项目反应函数中纳入随机题块效应参数 γ_n (类似题组模型中题组效应参数)来将迫选题块内题目的相互依赖性考虑到参数估计中，此模型即为 BRB-IRT。与 TIRT 相似，BRT-IRT 支持多种迫选题型，在用于 RANK-3 题型中时，同样需要进行二元编码(参考 TIRT 的编码方式)，因此其在 RANK-3 题型中采用的决策理论可被归类为瑟斯顿比较判断法则。那么个体 m 在题块 n 内选择题目 i 而非题目 j 的概率为：

$$P_{nm}\left(i > j \mid \theta_a, \theta_b\right) = \Phi_L\left(a_i\theta_a - a_j\theta_b - d_{ij} - \gamma_{nm}\right) = \frac{1}{1+\exp\left[-\left(a_i\theta_a - a_j\theta_b - d_{ij} - \gamma_{nm}\right)\right]} \tag{17}$$

其中题块 n 可以由 2 个及以上的测量不同维度的题目构成，与 MUPP-2PL 模型相似，d_{ij} 为截距参数 $\left(d_{ij} = a_i b_i - a_j b_j\right)$，$\gamma_{nm}$ 为个体 m 在题块 n 上的随机题块效应(random block effect)，其可被理解为题块 n 所测量的维度对个体作答的影响。不同题块会因所测维度的不同而产生不同的效应值。相似的，在传统题组模型中，题组效应是指一组题目的共同刺激(如阅读理解题目的篇章)对个体作答的影响。在参数估计上，BRB-IRT 与贝叶斯题组模型一致，采用 MCMC 方法。 H. Lee 和 Smith 从模拟研究中得到了与 TIRT 相似的对实践人员的建议，即需要采用混合计分型题块，才能获得比较可靠的参数估计结果，但他们仅模拟了 3 个维度的测验情境，高维情况下的表现还不得而知。另外，随机题块效应的大小并未对题目和能力参数的估计结果产生影响。

对于包含混合计分型题块会带来潜在的抗作假效力降低的这一争议问题，H. Lee 和 Smith 认为适合 BRB-IRT 的应用场景为低利害的作答情境，尤其是可以充分利用迫选测验能避免李克特式量表带来的其他作答反应偏差的这一优势，又不对抗作假有较高需求的场

景。如 2012 年 PISA(Programme for International Student Assessment)就在对学生的数学意向和学习策略量表上采用了迫选测验形式，通过控制潜在的由不同文化所带来的作答反应偏差来更好地了解学生的国际/跨文化差异。在 BRB-IRT 模型中，可以灵活地加入可能影响题目和特质分数的协变量，从而可以更好地分析人群间的差异。在公式(17)的基础上进行拓展，当包含影响所有特质的协变量(以性别变量为例)时，则为：

$$P_{nm}\left(i>j\,|\,\theta_a,\theta_b\right)=\frac{1}{1+\exp\left[-\left(a_i\theta_a-a_j\theta_b-d_{ij}-\gamma_{nm}+\beta gender_n\right)\right]} \tag{18}$$

当包含影响每个潜在特质的协变量时，为：

$$P_{nm}\left(i>j\,|\,\theta_a,\theta_b\right)=\frac{1}{1+\exp\left[-\left(a_i\theta_a-a_j\theta_b-d_{ij}-\gamma_{nm}+\left(\beta_a gender_n\right)-\left(\beta_b gender_n\right.\right.\right.} \tag{19}$$

通过公式(18)，可以解释性别是否对个体在 i 和 j 的选择上存在影响。通过公式(19)，可以解释性别是否对个体在与特质 a 和特质 b 有关的题目上的选择存在影响。

4　模型比较

4.1　模型构建思路

从实践的角度出发，通过已有的迫选 IRT 模型可以看到，迫选模型开发的一个方向是使其适合更多的题块组合方式，如 PICK、RANK 或 MOLE，另一个方向是使其适合不同反应模式的题目。依据题型和题目反应模型，已有迫选模型的总结见表 4。

表 4　模型总结

	PICK	RANK	MOLE
展开反应模型	MUPP-GGUM	GGUM-RANK	GGUM-RANK
优势反应模型	TIRT/MUPP-2PL/RIM/BRB IRT	TIRT/BRB IRT/ELIRT/GLIRT	TIRT/BRB IRT/ELIRT/GLIRT

TIRT、MUPP-2PL 和 BRB-IRT 均适合于优势反应模式题目，且选择了 2PLM 作为题目反应函数，只不过 MUPP-2PL 仅适合于 PICK-2 题型，其他两个均可通过对数据的二元编码应用于多种题型。TIRT 在应用于 PICK-2 题型时，在题块反应方程构建上与 MUPP-2PL 等价(Morillo et al., 2016)，只不过 TIRT 使用的是 Probit 链接函数，MUPP-2PL 使用的是 Logit 链接函数，而两模型在理论上的等价性在模拟研究中也得到了体现，两模型在大部分条件下的估计结果非常一致，除了 MUPP-2PL 对潜在特质及潜在特质之间关系的估计优于 TIRT 在

同等条件下的结果。另外 Morillo 等人的实证研究发现，MUPP-2PL 与 TIRT 无论是对题目参数还是对潜在特质的估计均具有极高的相似性(相关系数均在 0.9 附近)，这在一定程度也证明了 TIRT 所采用的的瑟斯顿比较判断法则与应用在 PICK-2 题型上的布拉德利-特里模型的内在等价性，只不过 TIRT 由于没有使用题目的先验信息导致估计结果总体偏极端化。为了将题块内题目之间的相互依赖性考虑到参数估计中，BRB-IRT 在 MUPP-2PL 的基础上加入了随机题块效应，而在 TIRT 中，则是通过构建题目间的协方差矩阵来实现的，H. Lee 和 Smith(2020a)的实证研究也表现出 BRB-IRT 与 TIRT 结果的高度一致性。

虽然 RIM 也为支持优势反应模式题目的迫选模型，但与 TIRT、MUPP-2PL 和 BRB-IRT 相比，它的题目反应函数为 Rasch 模型，在潜在特质分数原点的选择上也不一致。由于对θ意义的解释不同(RIM 将θ视为个体内部潜在特质的心理分化程度，而在其他三个模型中均被解释为真正意义的潜在特质的常模分数)，RIM 以个体内部均值或 0 为参照点/原点，因此只限制了个体内所有θ和为 0，并未对人群中的分布作假设，而其他模型均假设θ在人群中呈多元正态分布形态，θ的参照点为人群均值。显然，RIM 适合测量目的为寻找个体内部特质的排序状态的测验，TIRT、MUPP-2PL 和 BRB-IRT 适合目标在于比较不同个体之间的分数差异的测验。而 RIM 与传统计分方式得到的同为内部排序结果，那么相比传统计分，RIM 的优势性是否足以吸引实践人员转而采用更复杂的计分方法，需要更多研究去探索。

4.2　参数估计方法

在迫选模型的参数估计中，从估计内容上分为题目参数估计和潜在特质估计，从估计算法上主要分为传统估计算法和 MCMC 方法，从估计流程上主要分为联合估计和两步走策略。

在本文提及的 6 个模型中，仅 MUPP-GGUM 没有采用题目与能力参数的联合估计方法，如 3.4.1 中的流程所述，其是一种两步走策略：计算$P(i)$与$Q(i)$所需的题目参数是在第 2-3 步通过李克特式量表数据预先标定的，第 7 步基于 MUPP-GGUM 进行能力估计时使用了与题目标定时不同类型的迫选作答数据。因此此模型的一个隐含强假设是，题目参数具有跨测验形式的一致性。这种流程非常有利于题库的管理，进而方便迫选自适应测验的开发。在第 7 步对潜在特质进行估计时，Stark 等(2005,2012)采用一种近似牛顿迭代的 BFGS(Broyden-Fletcher-Goldfarb-Shanno)方法来实现高维能力估计的极大后验概率算法(Maximum A Posteriori, MAP)，BFGS 提供了一种近似梯度的数值计算方法使研究者可以免去 MAP 中所需要的黑森矩阵(Hessian Matrix)的推导，而高维情境下此矩阵的推导是非常繁琐的。Stark 使

用了 DFPMIN (Press et al., 1986)来实现 BFGS 算法，也可在 R 中通过指定 optim 函数中 method 参数为 L-BFGS-B 来实现。在题目参数标定这一环节，GGUM 最近几年在参数估计上也有了较多的突破(Roberts & Thompson, 2011)，并有相关的 R 包 GGUM (Tendeiro & Castro-Alvarez, 2018)、mirt (Chalmers, 2012)和 bmggum(Tu et al., 2021)支持。

TIRT 是基于结构方程模型开发的，且提出时间较长，现有多种成熟的软件(如 Mplus 等结构方程建模软件)和开源 R 包 thurstonianIRT (Bürkner, 2018)可用于其参数估计。为方便实践者使用，Brown 和 Maydeu-Olivares (2012)提供了输入测验设计就可以导出 Mplus 语句的 Excel 宏(http://annabrown.name/software)。而在 thurstonianIRT 包中，Bürkner 提供了数据模拟的函数，并作为一个接口供用户选择 lavaan 包(Yves Rosseel, 2012)，Mplus 或者 Stan(Stan Development Team, 2020)来作为模型拟合的内在处理方法，并可根据用户提供的信息自动生成三种方法的代码(Bürkner et al., 2019)。在 Mplus 或 lavaan 中，题目参数可使用未加权的最小二乘法或对角加权最小二乘法来估计。而 Stan 是一种概率编程语言，使用 MCMC 来拟合贝叶斯模型，因此 Bürkner 也提供了使 TIRT 能够采用 MCMC 估计的方便接口。潜在特质的估计可使用期望后验算法(Expected A Posteriori, EAP)或 MAP，EAP 适用于维度数量为 1-2 个时，MAP 适用于维度数量较多时(Brown, 2016)，因为维度数量升高时，会导致 EAP 中的数值积分的节点数呈指数级增长。

显然 TIRT 配套软件的开发为实践者提供了极大的便利性，这也是 TIRT 应用广泛的原因之一，但也存在一些质疑，如 Bürkner 等(2019)在使用 Mplus 和 Lavaan 拟合 TIRT 时发现有严重的模型无法收敛问题，特别是在大型测验的条件下(如 5 维度测验，每维度有 27 个题块，模型收敛率仅 0.3 左右)。除此之外还需要较高的运行内存(如 30 维度测验，每维度有 9 个题块，模型需要 32GB 的运行内存)，否则需要在代码中指定不计算卡方和标准误等拟合指标以减少运行时间和运行压力。最常见的报错是方差为负，通常需要指定维度间关系或因子载荷来促进收敛，但估计结果同样也会非常依赖这些固定值。使用 MCMC 方法拟合 TIRT 时没有不收敛和内存不够的问题，这得益于贝叶斯算法的自身优势。因此，考虑到 TIRT 在模型识别上的敏感性，如若在维度较高的测验中考虑使用 TIRT，需要在测验开发时就要充分保证题目的质量，如对题目进行单维性检验以保证题目的单维性特征。在选择估计方法时，需要考虑运行内存的问题。否则模型不收敛或因内存受限而无法获得任何估计结果会非常影响测验开发者对测验质量和模型的信心。最后从估计速度上来说，可能由于 Mplus 采用的未加权的最小二乘法是有限信息估计方法，在相同的测验条件下，其比 stan 的估计速度通常会快数倍，因此在非大型测验情境下，推荐先使用上述所举资源中 Mplus 进行分析。

与 TIRT 所采用的传统估计方法不同，后来模型的提出者均将参数估计算法落脚在了 MCMC 上。它是一种概率派、全信息的参数估计方法，不需要复杂的数学推导，仅需研究者构建合理的后验概率分布函数，并可以实现与频率派算法(极大似然估计等)相似的估计精度。MUPP-2PL、GGUM-RANK、RIM 和 BRB-IRT 模型均采用了 Metropolis-Hasting MCMC 算法，基于迫选数据进行题目和能力参数的联合估计。它们在先验信息的选择上并无明显区分，但它们所依托的估计软件有所不同。GGUM-RANK 使用的是 Ox(Doornik, 2009)，BRB-IRT 使用 OpenBUGS 3.2.3 (Lunn et al., 2009)，RIM 则使用 WinBUGS (Spiegelhalter et al., 2003) 或 JAGS (Plummer, 2003)，此外当维度数量少于 4 个时，RIM 推荐在 ConQuest(Adams et al., 2015)软件或者 R 包 TAM(Kiefer et al., 2016)中选择 MMLE 算法进行参数估计。在这些软件中，WinBUGS 和 OpenBUGS 相对比较慢，而 Bürkner 等(2019)针对 TIRT 开发的 MCMC 方法使用的是 Stan 语言，由于其采用了更先进的 NUTS(No-U-Turn sampler)或 HMC(Hamiltonian Monte Carlo)抽样方法，大幅提升了估计速度。在模型收敛的评价标准上，它们均采用了 Gelman 和 Rubin(1992)的 $\hat{R}$ 统计量(低于 1.2 则说明参数已收敛)。虽然这些模型均没有较大的收敛性问题，但需要实践人员比较深入地了解 MCMC 相关的知识和实施步骤，且 MCMC 方法的主要缺点是估计时间较长(Kim & Bolt 2007)，如 BRB-IRT 的一个模拟条件(1000 名被试，3 个维度，共 8 个 RANK-3 题块，重复次数为 25)需要长达 6 天的时间才能完成估计。针对各类模型参数估计方法的总结见表 5。

表 5　模型参数估计方法总结

参数估计方法	使用软件	优点	不足
两步走： 1.基于李克特式量表数据预标定题目参数 2.BFGS 估计能力	1. R 包：GGUM/mirt/bmggum 2. DFPMIN/R 包：stats	题目参数预先标定便于自适应题库管理	在迫选数据上使用李克特题目参数估计能力存在题目参数跨测验形式不一致的风险
加权的最小二乘法/对角加权最小二乘法	Mplus R 包：thurstonianIRT（Mplus/Lavaan 方法）	估计用时短，易用性强	高维情境下不易收敛，内存占用过高，有时需舍弃拟合指标的计算
MCMC	Ox/WinBUGS/JAGS/OpenBUGS R 包: thurstonianIRT（Stan 方法）	无收敛性问题	估计用时长，易用性不足

5 应用研究现状

迫选 IRT 模型被广泛应用于工业组织心理学领域，如 TIRT 在多种商业化测验中得到了应用，如 OPQ32r(Occupational Personality Questionnaire)和 CCSQ(Customer Contact Styles Questionn)两种性格测验(SHL, 2018; Brown & Maydeu-Olivares, 2011)，也被用于开发评估适应不良人格特征的测验(Assessment of Work-Related Maladaptive Personality Traits) (Guenole et al., 2016)。在 360 度反馈测验中也被证实使用迫选题型的测验并采用 TIRT 进行计分比使用传统李克特评分题目有更好的结构效度与聚合效度(Brown et al., 2017)。MUPP-GGUM 在员工人格自适应测验(the Adaptive Employee Personality Test, Adept-15)(Aon Hewitt, 2015)和美国军队选拔所开发的自适应人格测评工具 TAPAS(Tailored Adaptive Personality Assessment System) (Stark et al., 2014)上得以应用，这 2 个测验也是迫选模型在计算机化自适应方向的突破性尝试。同时，题目参数不变性检验作为测验开发流程的一个重要环节，在迫选模型上的检验方法也正逐步被开发与完善。在实践人员更加关注的效度研究领域，也积累了相当多的证据。因此，本文将对迫选模型在参数不变性检验、计算机化自适应测验和效度研究 3 个方面进行迫选模型应用研究的现状分析。

5.1 参数不变性检验

通常，测验开发者需要对题目参数的不变性进行检验(也即测量一致性检验)，以保证所有作答者对题目的理解或者题目所表达的内涵是相同的。在迫选测验情境下，题目参数不变性可根据不变性情境分为 2 个具体问题：跨题块一致性和跨人群一致性。不具有参数不变性的题目意味着其作答概率会受到除测量目标外其他因素的影响。

跨题块一致性是指同一题目在与不同题目组合为题块时，其是否具有参数不变性，如有题块 1{A,B,C}和题块 2{A,D,E}，它们的共同题目为 A 描述，如果 A 的题目参数在两题块的估计结果差异不大，则说明参数没有受到其他题目的影响，具有跨题块参数不变性。Lin 和 Brown (2017)基于 TIRT 模型，比较了 RANK-3 和 MOLE-4 两种题型的两套迫选测验的参数不变性，后者仅在前者的每个题块上新增了一道题目，所以每对题块之间的共同题比例为 75%，结果发现仅有少量题目存在较大偏差。

跨人群一致性是指一道题目在来自不同背景的人群组(如不同性别、不同文化背景、不

同测验情境的人群)之间是否具有参数不变性，而对此不变性的检验也称之为题目功能性差异检验(Differential Item Functioning, DIF)，如果题目参数在不同组之间发生了较大改变，就意味着此题目的作答概率会受到个体背景的影响，如果测验中包含较多此类题目将会降低测验效度，也有失公平性。在开发迫选测验时，首先需确保单题题库具有良好的测量学指标，如可接受的区分度指标、没有 DIF 等 (Stark et al., 2005; SHL, 2018)，这些题目质量分析通常采用李克特等单一刺激量表形式进行，但当题目组合为题块时，则可能产生新的 DIF(不同组别的人群因为题目情境发生改变而产生与单一刺激题目不同的反应偏好)。因此，基于迫选数据进行 DIF 检验是势在必行的。

H. Lee 和 Smith(2020b)基于多组 CFA(multiple group confirmatory factor analyses)框架提出了通过模型的整体拟合指数差异来检验 TIRT 测量不变性(Measurement Invariance, MI)的分析方法，并建议将 $\Delta CFI > 0.007$ 和 $\Delta CFI > 0.001$ 分别作为尺度非一致性(metric non-invariance)和标量非一致性(scalar non-invariance)的临界值，但此方法无法具体到题目来进行筛查，而题目层面的参数非一致性即为 DIF。P. Lee 等(2020)则针对 TIRT 模型的区分度和截距指标提出了一种综合 Wald 检验 TIRT DIF 方法(omnibus Wald tests)，并通过模拟研究证明在自由基线(free baseline)方法下检出效率较高：随着样本量和 DIF 量的增加，检出率接近 1，I 型错误率接近 0.05。Qiu 和 Wang(2021) 提出了 3 种 RIM 的 DIF 检验方法，EMD (equal-mean-difficulty), AOS(all-other-statement)和 CS(constant-statement) 方法，最终通过模拟研究发现 CS 方法在测验含有 DIF 题目时的表现优于其他两种方法。

5.2 计算机化自适应测验

由于人类性格特点的复杂性，性格测评工具的测量维度也通常是高维的，如 OPQ32r 测量了 32 个性格维度。维度越多，意味着所需要的题目也越多，测验总长度就会达到惊人的程度。从个体感受而言，题量过长会使个体疲惫度增高进而对测验感到厌烦导致粗心作答，特别是在招聘情境下使用时，甚至会对应聘企业或测评提供方产生不好的印象。从测评效率来说，当个体在某些维度上通过少量题目已经达到可接受的测评精度时，即可以对个体在这些维度上有比较确定的判断，在后续集中投放对其评价不确定性更高的维度的题目，就能尽快使对个体所有维度的评估都能达到一个可靠的程度，从而提升测评效率，降低企业招聘在测评上花费的时间和成本。而解决以上问题的思路之一就是开发 CAT 版本的迫选测验。

早在 15 年前，迫选 CAT 测验就已经在美国海军人员选拔中得到了应用，该测验由

Houston 等人(2006)开发，全称为美国海军自适应人格量表(Navy Computer Adaptive Personality Scales, NCAPS)，共测量了 19 个性格维度。在进行自适应评估时，会依据个体当前能力抽取同一维度下处于两端的题目并参考其称许性水平进行配对，因此其为单维 PICK-2 题型的自适应测验。Stark 等人(2012)在 MUPP-GGUM 的基础上提出了适用于多重-单维 PICK-2 题型(可使用单维和多维题块)的共 6 个步骤的迫选自适应流程，与传统 CAT 最大的区别是需要考虑单维题块的比例和预先遍历并存储多维题块的维度组合形式。如对于一个 3 维度的测验，它的维度组合形式有 1-1、2-2、3-3、1-2、1-3、2-3，并在此基础上控制内容的平衡。为了加速对特质水平的估计，该流程推荐使用环形维度链接策略(Circular Dimensional-Linking)，即使用最少的题块链接所有的维度，如一个 5 维度测验可使用维度组合 1-2、2-3、3-4、4-5、5-1。以上两个研究均证明了 CAT 相比非 CAT 对效率的提升是非常明显的，迫选 CAT 只需要非自适应测验一半的题目就能达到同样的准确性。另外，TAPAS 也是为美国军队选拔所开发的自适应人格测评工具，同样基于 MUPP-GGUM (Stark et al., 2014)。

除以上提及的配对题型的迫选 CAT 测验外，近期 Joo 等人(2020)基于 GGUM-RANK 提出了适用于 RANK 题型的迫选 CAT 方法，并通过模拟研究指出单维题块似乎不是必须的，而在 Stark 等人(2012)的研究中推荐加入总题量 5%的单维题块。

Chen 等(2020) 提出了 3 种子库选题策略(subpool selection strategies)来提升选题效率和控制题目曝光率，这三种策略分别为序列策略(The Sequential Strategy)、多项式策略(The Multinomial Strategy)和高 SE 策略(The High-SE Strategy)。序列策略与 Stark 等人(2012)的第一步相似，需要先构建所有维度的组合形式，如 6 维度的 RANK-3 测验，将有 20 种维度组合数，每个组合内由各维度的题目形成一个子题库。接下来，将从各个组合题库内依据信息量开始循环抽题，直至达到终止标准。由于每个被试都按这个顺序进行抽题，会有题目组合形式曝光的风险，但更值得担心的是当维度数量过多时，如 12 个维度时会产生 220 个子题库，如果最大测验长度被限制为 50 题，将会导致 170 个子题库的题目不被抽到。多项式策略通过根据多项式分布随机选择子题库来解决序列策略的问题。当确定子题库后，根据目标测验长度 L 和预设的在每个子题库所抽取题块数 T ，给出每个子题库的抽取概率 $P_{target} = T / L$ 。例如：当测验长度为 100 且预设子题库 1 抽取 10 题时， $T[1] = 10$ ，子题库 1 被抽取到的概率为 0.1。根据概率进行随机抽取题库，当某个子题库已抽取的题块数量达到其在 T 中的预设值后，剔除该题库后重新计算概率，再进行后续抽取。高 SE 策略则是先判定个体在哪些维度上具有最高的 SE，再选择对应维度组合的子题库题块。与全题库抽

取策略下用时(6.72s)相比，子题库抽取策略下用时均有下降，全部在 1s 以内，且测量精度没有明显下降。但为了达到相似测量精度，序列策略所用的题块数量要高于其他策略，同时，高 SE 策略在内容平衡方面表现较差，多项式策略综合表现更优。

此外，针对题目曝光率的控制，Chen 等(2020)提出了 RSHO(revised Sympson-Hetter online)方法。在进行题块选择时，先根据信息量确定最适合题块，计算该题块中每个题目在备选题库中和已作答题块中的数量，分别除以已作答题块数形成 $P(S)$ 和 $P(A)$。再将 $P(S)$ 和 $P(A)$ 与题目最大曝光率 r 作对比，形成该题目描述的 pks， pks 具体计算方式如下：

$$pks=\begin{cases}0, & P(A)\ge r\\ \dfrac{r}{P(S)}, & P(A)\le r \text{ and } P(S)\ge r\\ 1, & P(A)\le r \text{ and } P(S)\le r\end{cases} \tag{20}$$

对于一个题块，会形成多个 pks，选取题块内数值最小的 pks 形成题块的 pk 值，再生成一个 0 到 1 之间的随机数字与 pk 进行对比，若随机数字小于 pk 则抽题，否则剔除该题块后重复以上步骤。$P(S)$ 和 $P(A)$ 的初始值设置为 0，且在每次选题时均需要重新计算 pk 值。RSHO 方法在稍微牺牲测量精度的前提下控制了题目的曝光率。

就迫选 CAT 测验的重测信度这一问题，Seybert 和 Becker(2019)指出题目不一致带来的误差降低了 CAT 测验的重测信度，因为在测验施测过程中受到作答者能力、选题策略等多方面影响，很难找到完全相同的两份 CAT 试卷，所以 CAT 的重测信度更像是传统测验的复本重测信度(在不同时间点对个体施测 2 份复本测验)。其研究表明迫选 CAT 测验的重测信度低于传统李克特式量表，但与传统李克特式量表的复本重测信度相当。

5.3 效度研究

由迫选 IRT 模型获得的潜在特质分数是否能反应出个体的真实特点，为了回答这一问题，研究者主要从 4 个方向进行了探索。首先是探索迫选 IRT 计分是不是比传统计分对潜在特质及其之间关系的恢复性更好(Hontangas et al., 2015; Hontangas et al., 2016; Oswald et al., 2015)。相比传统计分，使用迫选 IRT 模型进行特质分数的估计能带来显著的测量精度的提升，这几乎是这个方向所有研究共同的结论，也给了研究者极大的信心去开发更多的迫选 IRT 模型，但也有 Wang 等(2017)对模型中潜在特质 θ 不一样的解读思路和 Schulte 等(2021)

指出并非在所有情境下采用 TIRT 模型都能利用到 IRT 的优良性质，甚至在高维情境下得到的分数依然是部分自模的。此外还有 Walton 等(2020)的研究指出 TIRT 模型在大五人格量表上区分效度不如传统计分下自模式数据。那么从这些模型中拿到的分数能在多大程度上被理解为传统的常模性分数还值得更多的研究去探索，因为这直接关系到这些分数是否能够像常模性分数那样做人员选拔的应用或与外部效标进行关联性分析。而第二个方向则试图通过探索迫选 IRT 与李克特式单一刺激量表得到的潜在特质分数之间的关系(Zhang et al., 2020; Watrin et al., 2019; Joubert et al., 2015; Guenole et al., 2016)来回答上述问题，在这些研究中单一刺激量表分数被认为最符合个体潜在特征的真值，如果通过迫选模型得到的分数与其在分数原点、尺度和维度关系上均能保持较高的相似性也就证明了二者的等价性，那么就能像使用李克特式量表结果那样来对迫选模型得到的结果做相关的分析了。第三个方向是探索迫选测验的抗作假能力。当匹配了迫选题块内的社会称许性时，迫选测验的抗作假能力要优于李克特式量表(Wetzel et al., 2020)。与用 TIRT 分析迫选测验相比，利用等级反应模型(Graded Response Model, GRM)分析李克特式量表无法有效区分高能力者，因为作答者倾向表现得更好，导致那些能够体现个体高能力的题目的区分度较低(Dueber et al., 2018)。第四个方向则是探索迫选 IRT 模型在非自评情境下的使用。因为他评李克特式量表同样存在共同方法偏差，不同评估人的评价受其内在理想行为标准的影响，导致评分者一致性信度较低。当在 360 度评估中应用迫选 IRT 模型时，相比李克特式量表，不同层级评估人的评分者一致性信度上升，题目的结构效度也更好(Brown et al., 2017)。

6 研究展望

迫选测验作为一种能有效抵抗作假、作答反应偏差且能提升个体作答效率的测评形式，迫选 IRT 模型的研究依然具有很大的潜力，尤其是在非认知类、高利害情境测评的应用上。结合已有研究未解决的问题提出以下几个对未来研究的展望方向：模型拓展研究、题目参数不变性研究、迫选 CAT 研究和效度研究。

6.1 模型拓展研究

目前已有的迫选模型均适用于常规题型，如 PICK-2，RANK-3。未来还可以探索这些模型是否可以通过对数据的重新编码来支持 Q 分类题型。另外，还存在 PICK-2 题型的变体形

式，如 Adept-15 测验(Aon Hewitt, 2015)，即在让候选人选择最符合自己的一项时，同时给出选择此项的意愿程度(见表 6)，因此可称之为 PICK-2 的多级计分形式。

表 6　PICK-2 多级计分

	比较符合	非常符合
A 寻找事物的不足		√
B 探索陌生的领域		

这种题型细化了个体的选择行为，理论上提供了更多信息量，从原来的 2 个计分点，扩充至了 4 个，但也增加了题目的认知负荷，因此似乎只能在较小题块中使用。还未发现有直接的模型来拟合此种题型的数据，多级计分版本的迫选模型的研究有待探索。

6.2　基于各模型的参数不变性研究

针对参数的跨题块一致性问题，延续 Lin 和 Brown (2017)针对 TIRT 的研究，当共同题比例降低时，是否还能有较高比例的题目的参数具有跨题块不变性还有待研究。另外针对其他模型的跨题块一致性的研究有待展开。

目前仅有针对 TIRT(H. Lee & Smith, 2020b; P. Lee et al., 2020)和 RIM(Qiu & Wang, 2021)的参数不变性检验的研究，未来研究除了需要开发其他迫选模型 DIF 检验方法，也需丰富或提升已有的 DIF 检验方法，使之对多种来源的 DIF 更加敏锐。

6.3　迫选 CAT 研究

虽然迫选 CAT 在实证研究上积累了较多的经验，但已开发的自适应流程在进行潜在特质估计时所采用的题目参数均为通过单一刺激量表数据预先标定的，所使用的题库均为单题题库，并非题块库，在进行题目选择时将即时进行题目的组合形成迫选题块，那么题目的跨题块一致性在这种 CAT 流程下对潜在特质估计的影响需要进一步研究。另外高维情境下题块维度组合形式和测验长度均会大幅增加，这对内容平衡和测验效率带来了挑战，未来可进一步探索如何在高维情境下发挥 CAT 的优势。虽然 Chen 等(2020)提出的子题库选题策略和题目曝光控制方法不涉及计分相关内容，可以拓展至基于其他非 RIM 模型构建的 CAT 测验中，但具体表现如何还需要研究去探索。另外，多项式策略等控制方法无法直接应用于变长测验，未来可进一步探索在变长测验中如何构建更合适的选题策略。

6.4 效度研究

目前有较多研究集中于对比迫选测验与李克特式量表在针对同一测量内容时是否产出了相似的结果，以此来证明其抗作假效力和常模性分数的恢复程度，但二者在测验形式上的区分性和李克特题型所带来的作答反应偏差必然会引入一些误差，未来如何最大限度控制这些偏差或者是否存在更好的效度研究思路值得探索。在迫选形式上，题块越大，抵抗作假能力越强，但也增加了认知负荷(Wetzel et al., 2020)，在未来研究中可以探索抗作假效力和认知负荷在题块大小上的平衡点。另外，已有的效度研究大多数围绕 TIRT 展开，GGUM-RANK 等新模型的效度研究有待探索。

参考文献

李辉，肖悦，刘红云. (2017). 抗作假人格迫选测验中瑟斯顿 IRT 模型的影响因素. *北京师范大学学报(自然科学版)*, *053*(005), 624–630.

连旭，卞迁，曾劭婵，车宏生. (2014). MAP 职业性格迫选测验基于瑟斯顿 IRT 模型的拟合分析[摘要]. *全国心理学学术会议*, 中国北京.

骆方，张厚粲. (2007). 人格测验中作假的控制方法. *心理学探新*, *27*(4), 78-82.

王珊，骆方，刘红云. (2014). 迫选式人格测验的传统计分与 IRT 计分模型. *心理科学进展* *22* (3), 9.

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). *ACER Conquest 4.0 [Computer program]*. Melbourne: Australian Council for Educational Research

Aon, Hewitt. (2015). *2015 Trends in global employee engagement report*. Lincolnshire, IL: Aon Corp.

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology*, *69*(1), 49–56.

Bartram, D. (1996). The relationship between ipsatized and normative measures of personality. *Journal of Occupational and Organizational Psychology*, *69*(1), 25–39.

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment*, *15*(3), 263–272.

Block, J. (1963). The Q-sort method in personality assessment and psychiatric research. *Archives of General Psychiatry*, *136*(3), 230–231.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3/4).

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*(2), 153–168.

Brown, A. (2016). Item response models for forced-choice questionnaires: A common framework. *Psychometrika*, *81*(1), 135–160.

Brown, A., Inceoglu, I., & Lin, Y. (2017). Preventing rater biases in 360-degree feedback by forcing choice. *Organizational Research Methods*, *20*(1), 121–148.

Brown, A., & Maydeu-Olivares, A. (2010). Issues that should not be overlooked in the dominance versus ideal point controversy. *Industrial and Organizational Psychology*, *3*(4), 489–493.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*(3), 460–502.

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods*, *44*(4), 1135–1147.

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, *18*(1), 36–52.

Bürkner, P.-C. (2018). thurstonianIRT: Thurstonian IRT models in R. *Journal of Open Source Software*, *4*(42), 1662.

Bürkner, P.-C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of Thurstonian IRT models. *Educational and Psychological Measurement*, *79*(5), 827–854.

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *The Journal of Applied Psychology*, *104*(11), 1347–1368.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29.

Chen, C.-W., Wang, W.-C., Chiu, M. M., & Ro, S. (2020). Item selection and exposure control methods for computerized adaptive testing with multidimensional ranking items. *Journal of Educational Measurement*, *57*(2), 343–369.

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*(4), 523–562.

Clemans, W. V. (1966). An analytical and empirical examination of some properties of ipsative measures. *Psychometric Monographs*, 14.

Closs, S. J. (1996). On the factoring and interpretation of ipsative data. *Journal of Occupational and Organizational Psychology*, *69*(1), 41-47.

Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychological Review*, *57*(3), 145–158.

Doornik, J. A. (2009). An object-oriented matrix programming language Ox 6. London, England: Timberlake Consultants.

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology*, *3*(4), 465-476.

Dueber, D. M., Love, A. M. A., Toland, M. D., & Turner, T. A. (2018). Comparison of single-response format and forced-choice format instruments using Thurstonian item response theory. *Educational and Psychological Measurement*, *79*(1), 108–128.

Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance*, *16*(1), 1–23.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457-472.

Guenole, N., Brown, A., & Cooper, A. (2016). Forced-choice assessment of work-related maladaptive personality traits: Preliminary evidence from an application of Thurstonian item response modeling. *Assessment*, *25*(4), 513–526.

Gwet, K. L. (2014). *Handbook of inter-rater reliability (4th ed.): The defnitive guide to measuring the extent of agreement among raters*. Gaithersburg, MD: Advanced Analytics, LLC.

Hendy, N., Krammer, G., Schermer, J. A., & Biderman, M. D. (2021). Using bifactor models to identify faking on Big Five questionnaires. *International Journal of Selection and Assessment*, *29*(1), 81–99.

Hontangas, P. M., La Torre, J. de, Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing traditional and IRT scoring of forced-choice tests. *Applied Psychological Measurement*, *39*(8), 598–612.

Hontangas, P. M., Leenen, I., La Torre, J. de, Ponsoda, V., Morillo, D., & Abad, F. J. (2016). Traditional scores versus IRT estimates on forced-choice tests based on a dominance model. *Psicothema*, *28*(1), 76–82.

Houston, J., Borman, W., Farmer, W., & Bearden, R. (2006). *Development of the Navy Computer Adaptive Personality Scales (NCAPS)* (NPRST-TR-06-2). Millington, TN: Navy Personnel Research, Studies, and Technology.

Huang, J., & Mead, A. D. (2014). Effect of personality item writing on psychometric properties of ideal-point and likert scales. *Psychological Assessment*, *26*(4), 1162–1172.

Hurtz, G., & Donovan, J. (2000). Personality and job performance: The Big Five revisited. *The Journal of Applied Psychology*, *85*(6), 869–879.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance*, *13*(4), 371–388.

Joo, S.-H., Lee, P., & Stark, S. (2018). Development of information functions and indices for the GGUM-RANK multidimensional forced choice IRT model. *Journal Of Educational Measurement*, *55*(3), 357-372.

Joo, S.-H., Lee, P., & Stark, S. (2020). Adaptive testing with the GGUM-RANK multidimensional forced choice model: Comparison of pair, triplet, and tetrad scoring. *Behavior Research Methods*, *52*(2), 761–772.

Joubert, T., Inceoglu, I., Bartram, D., Dowdeswell, K., & Lin, Y. (2015). A comparison of the psychometric properties of the forced choice and likert scale versions of a personality instrument. *International Journal of Selection and Assessment*, *23*(1), 92–97.

Kiefer, T., Robitzsch, A., & Wu, M. (2016). TAM: Test analysis modules (R package version 1.995-0) [Computer program]. Retrieved from https://cran.r-project.org/web/packages/TAM/index.html

Kim, J.-S., & Bolt, D. (2007). Estimating item response theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practice*, *26*(4), 38–51.

Lee, H., & Smith, W. Z. (2020a). A Bayesian random block item response theory model for forced-choice formats. *Educational and Psychological Measurement*, *80*(3), 578–603.

Lee, H., & Smith, W. Z. (2020b). Fit indices for measurement invariance tests in the Thurstonian IRT model. *Applied Psychological Measurement*, *44*(4), 282–295.

Lee, P., Joo, S.-H., & Stark, S. (2020). Detecting DIF in multidimensional forced choice measures using the Thurstonian item response theory model. *Organizational Research Methods*, *24*(4), 739–771.

Lee, P., Joo, S.-H., Stark, S., & Chernyshenko, O. S. (2019). GGUM-RANK statement and person parameter estimation with multidimensional forced choice triplets. *Applied Psychological Measurement*, *43*(3), 226–240.

Li, M., Sun, T., & Zhang, B. (2021). autoFC: An R package for automatic item pairing in forced-choice test construction. *Applied Psychological Measurement,* Advance online publication. https://doi.org/10.1177/01466216211051726.

Lin, Y., & Brown, A. (2017). Influence of context on item parameters in forced-choice personality assessments. *Educational and Psychological Measurement*, *77*(3), 389–414.

Luce, & Duncan, R. (1959). On the possible psychophysical laws. *Psychological Review*, *66*(2), 81–95.

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology*, *15*(3), 215–233.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*(25), 3049-3067.

Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research*, *45*(6), 935–974.

Morillo, D., Leenen, I., Abad, F. J., Hontangas, P., La Torre, J. de, & Ponsoda, V. (2016). A dominance variant under the multi-unidimensional pairwise-preference framework: Model formulation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *40*(7), 500–516.

Oswald, F. L., & Schell, K. L. (2010). Developing and scaling personality measures: Thurstone was right—But so far, Likert was not wrong. *Industrial and Organizational Psychology*, *3*(4), 481–484.

Oswald, F. L., Shaw, A., & Farmer, W. L. (2015). Comparing simple scoring with IRT scoring of personality measures: The Navy Computer Adaptive Personality Scales. *Applied Psychological Measurement*, *39*(2), 144–154.

Pavlov, G., Shi, D., Maydeu-Olivares, A., & Fairchild, A. (2021). Item desirability matching in forced-choice test construction. *Personality and Individual Differences*, 183, 111114.

Plummer, M. (2003, March). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Paper presented at the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1986). *Numerical recipes: The art of scientific computing.* New York: Cambridge University Press.

Qiu, X.-L., & Wang, W.-C. (2021). Assessment of differential statement functioning in ipsative tests with multidimensional forced-choice items. *Applied Psychological Measurement*, *45*(2), 79–94.

R Core Team. (2021): *R: A language and environment for statistical computing.* vienna, Austria. Available online at https://www.R-project.org/.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, *24*(1), 3–32.

Roberts, J. S., & Thompson, V. M. (2011). Marginal maximum a posteriori item parameter estimation for the generalized graded unfolding model. *Applied Psychological Measurement*, *35*(4), 259–279.

Sass, R., Frick, S., Reips, U.-D., & Wetzel, E. (2020). Taking the test taker's perspective: Response process and test motivation in multidimensional forced-choice versus rating scale instruments. *Assessment*, *27*(3), 572–584.

Saville, P., & Willson, E. (1991). The reliability and validity of normative and ipsative approaches in the measurement of personality. *Journal Of occupational Psychology*, *64*(3), 219–238.

Schulte, N., Holling, H., & Bürkner, P.-C. (2021). Can high-dimensional questionnaires resolve the ipsativity issue of forced-choice response formats? *Educational and Psychological Measurement*, *81*(2), 262–289.

Seybert, J., & Becker, D. (2019). Examination of the test–retest reliability of a forced-choice personality measure. *ETS Research Report Series*, *2019*(1), 1-17.

SHL. (2018). *OPQ32r technical manual*. SHL.

Sitser, T., van der Linden, D., & Born, M. P. (2013). Predicting sales performance criteria with personality measures: The use of the general factor of personality, the Big Five and narrow traits. *Human Performance*, *26*(2), 126–149.

Spiegelhalter, D., Thomas, A., & Best, N. (2003). WinBUGS version 1.4 [Computer program]. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

Stan Development Team. (2020). *RStan: the R interface to Stan*. http://mc-stan.org/

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement*, *29*(3), 184–203.

Stark, S., Chernyshenko, O. S., Drasgow, F., Nye, C. D., White, L. A., Heffner, T., & Farmer, W. L. (2014). From ABLE to TAPAS: A new generation of personality tests to support military selection and classification decisions. *Military Psychology*, *26*(3), 153–164.

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items. *Organizational Research Methods*, *15*(3), 463–487.

Stark, S. E. (2002). *A new IRT approach to test construction and scoring designed to reduce the effects of faking in personality assessment: The generalized graded unfolding model for multi -unidimensional paired comparison responses* (Unpublished doctorial dissertation). University of Illinois at Urbana-Champaign.

Tay, L., Ali, U. S., Drasgow, F., & Williams, B. (2011). Fitting IRT models to dichotomous and polytomous data: assessing the relative model–data fit of ideal point and dominance models. *Applied Psychological Measurement*, *35*(4), 280–295.

Tendeiro, J. N., & Castro-Alvarez, S. (2018). GGUM: An R package for fitting the generalized graded unfolding model. *Applied Psychological Measurement*, *43*(2), 172–173.

Thurstone, L.L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273–286.

Tu, N., Zhang, B., Angrave, L., & Sun, T. (2021). bmggum : An R package for Bayesian estimation of the multidimensional generalized graded unfolding model with covariates. *Applied Psychological Measurement, 45*(7–8), 553–555 .

Usami, S., Sakamoto, A., Naito, J., & Abe, Y. (2016). Developing pairwise preference-based personality test and experimental investigation of its resistance to faking effect by item response model. *International Journal of Testing*, *16*(4), 288–309.

Walton, K. E., Cherkasova, L., & Roberts, R. D. (2020). On the validity of forced choice scores derived from the Thurstonian item response theory model. *Assessment*, *27*(4), 706–718.

Wang, W.-C., Qiu, X.-L., Chen, C.-W., & Ro, S. (2016). Item response theory models for multidimensional ranking items. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative psychology research* (pp. 49-65). New York, NY: Springer.

Wang, W.-C., Qiu, X.-L., Chen, C.-W., Ro, S., & Jin, K.-Y. (2017). Item response theory models for ipsative tests with multidimensional pairwise comparison items. *Applied Psychological Measurement*, *41*(8), 600–613.

Watrin, L., Geiger, M., Spengler, M., & Wilhelm, O. (2019). Forced-choice versus Likert responses on an occupational Big Five questionnaire. *Journal of Individual Differences*, *40*(3), 134–148.

Wetzel, E., Frick, S., & Brown, A. (2020). Does multidimensional forced-choice prevent faking? Comparing the susceptibility of the multidimensional forced-choice format and the rating scale format to faking. *Psychological Assessment*, *33*.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Zhang, B., Sun, T., Drasgow, F., Chernyshenko, O. S., Nye, C. D., Stark, S., & White, L. A. (2020). Though forced, still valid: Psychometric equivalence of forced-choice and single-statement measures. *Organizational Research Methods*, *23*(3), 569–590.

Ziegler, M., MacCann, C., & Roberts, R. D. (Eds.). (2012). *New perspectives on faking in personality assessment*. Oxford, UK: Oxford University Press.

IRT-based scoring methods for multidimensional forced choice tests

LIU Juan, ZHENG Chanjin, LI Yunchuan, LIAN Xu

(Beijing Insight Online Management Consulting Co .,Ltd., Beijing 100102, China) (East China Normal University, Shanghai 200062, China)

Abstract: Forced-choice (FC) test is widely used in non-cognitive tests because it can control the response bias caused by the traditional Likert method, while traditional scoring of forced-choice test produces ipsative data that has been criticized for being unsuitable for inter-individual comparisons. In recent years, the development of multiple forced-choice IRT models that allow researchers to obtain normative information from forced-choice test has re-ignited the interest of researchers and practitioners in forced-choice IRT models. First, the six prevailing forced-choice IRT models are classified and introduced according to the adopted decision models and item response models. Then, the models are compared and summarized from two perspectives: model construction ideology and parameter estimation methods. Next, it reviews the applied research of the model in three aspects: parameter invariance testing, computerized adaptive testing (CAT) and validity study. Finally, it is suggested that future research can move forward in four directions: model expansion, parameter invariance testing, forced-choice CAT, and validity research.

Key words: forced choice test, ipsative data, TIRT, MUPP, GGUM-RANK